Functional Bregman Divergence and Bayesian Estimation of Distributions

B. A. Frigyik, S. Srivastava, and M. R. Gupta

***Abstract*— A class of distortions termed functional Bregman divergences is defined, which includes squared error and relative entropy. A functional Bregman divergence acts on functions or distributions, and generalizes the standard Bregman divergence for vectors and a previous pointwise Bregman divergence that was defined for functions. A recently published result showed that the mean minimizes the expected Bregman divergence. The new functional definition enables the extension of this result to the continuous case to show that the mean minimizes the expected functional Bregman divergence over a set of functions or distributions. It is shown how this theorem applies to the Bayesian estimation of distributions. Estimation of the uniform distribution from independent and identically drawn samples is used as a case study.**

***Index Terms*— Bregman divergence, Bayesian estimation, uniform distribution, learning**

BREGMAN divergences are a useful set of distortion functions that include squared error, relative entropy, logistic loss, Mahalanobis distance, and the Itakura-Saito function. Bregman divergences are popular in statistical estimation and information theory. Analysis using the concept of Bregman divergences has played a key role in recent advances in statistical learning [1]–[10], clustering [11], [12], inverse problems [13], maximum entropy estimation [14], and the applicability of the data processing theorem [15]. Recently, it was discovered that the mean is the minimizer of the expected Bregman divergence for a set of d-dimensional points [11], [16].

In this paper we define a functional Bregman divergence that applies to functions and distributions, and we show that this new definition is equivalent to Bregman divergence applied to vectors. The functional definition generalizes a pointwise Bregman divergence that has been previously defined for measurable functions [7], [17], and thus extends the class of distortion functions that are Bregman divergences; see Section I-A.2 for an example. Most importantly, the functional definition enables one to solve functional minimization problems using standard methods from the calculus of variations; we extend the recent result on the expectation of vector Bregman divergence [11], [16] to show that the mean minimizes the expected Bregman divergence for a set of functions or distributions. We show how this theorem links to Bayesian estimation of distributions. For distributions from the exponential family distributions, many popular divergences, such as relative entropy, can be expressed as a (different) Bregman divergence on the exponential distribution parameters. The functional Bregman definition enables stronger results and a more general application.

In Section 1 we state a functional definition of the Bregman divergence and give examples for total squared difference, relative entropy, and squared bias. In later subsections, the relationship between the functional definition and previous Bregman definitions is established, and properties are noted. Then in Section 2 we present the main theorem: that the expectation of a set of functions minimizes the expected Bregman divergence. We discuss the application of this theorem to Bayesian estimation, and as a case study compare different estimates for the uniform distribution given independent and identically drawn samples. Proofs are in the appendix. Readers who are not familiar with functional derivatives may find helpful our short introduction to functional derivatives [18] or the text by Gelfand and Fomin [19].

B. A. Frigyik is with the Department of Mathematics, Purdue University, West Lafayette, IN 47907 (e-mail: bfrigyik@math.purdue.edu).

S. Srivastava is with the Department of Applied Mathematics, University of Washington, Seattle, WA 98195 (e-mail: santosh@amath.washington.edu).

M. R. Gupta is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 (e-mail: gupta@ee.washington.edu). This author's work was supported in part by the United States Office of Naval Research.

I. Functional Bregman Divergence

Let $\left(\mathbb{R}^d, \Omega, \nu\right)$ be a measure space, where ν is a Borel measure and d is a positive integer. Let ϕ be a real functional over the normed space $L^p(\nu)$ for $1 \leq p \leq \infty$. Recall that the bounded linear functional $\delta\phi[f;\cdot]$ is the Fréchet derivative of ϕ at $f \in L^p(\nu)$ if

$$\begin{aligned}\phi[f+a] - \phi[f] &= \triangle\phi[f;a] \\ &= \delta\phi[f;a] + \epsilon[f,a]\,\|a\|_{L^p(\nu)} \end{aligned} \tag{1}$$

for all $a \in L^p(\nu)$, with $\epsilon[f,a] \to 0$ as $\|a\|_{L^p(\nu)} \to 0$ [19]. Then given an appropriate functional ϕ, a functional Bregman divergence can be defined:

Definition I.1 (Functional Definition of Bregman Divergence)**.** *Let $\phi : L^p(\nu) \to \mathbb{R}$ be a strictly convex, twice-continuously Fréchet-differentiable functional. The Bregman divergence $d_\phi : L^p(\nu) \times L^p(\nu) \to [0,\infty)$ is defined for all admissible $f, g \in L^p(\nu)$ as*

$$d_\phi[f,g] = \phi[f] - \phi[g] - \delta\phi[g; f-g], \tag{2}$$

where $\delta\phi[g;\cdot]$ is the Fréchet derivative of ϕ at g.

Here, we have used the Fréchet derivative, but the definition (and results in this paper) can be easily extended using other definitions of functional derivatives; a sample extension is given in Section I-A.3.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Functional Bregman Divergence and Bayesian Estimation of Distributions	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Electrical Engineering,Seattle,WA,98195**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A class of distortions termed functional Bregman divergences is defined, which includes squared error and relative entropy. A functional Bregman divergence acts on functions or distributions, and generalizes the standard Bregman divergence for vectors and a previous pointwise Bregman divergence that was defined for functions. A recently published result showed that the mean minimizes the expected Bregman divergence. The new functional definition enables the extension of this result to the continuous case to show that the mean minimizes the expected functional Bregman divergence over a set of functions or distributions. It is shown how this theorem applies to the Bayesian estimation of distributions. Estimation of the uniform distribution from independent and identically drawn samples is used as a case study.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

A. Examples

Different choices of the functional ϕ lead to different Bregman divergences. Illustrative examples are given for squared error, squared bias, and relative entropy. Functionals for other Bregman divergences can be derived based on these examples, from the example functions for the discrete case given in Table 1 of [16], and from the fact that ϕ is a strictly convex functional if it has the form $\phi(g) = \int \tilde{\phi}(g(t))dt$ where $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$, $\tilde{\phi}$ is strictly convex and g is in some well-defined vector space of functions [20].

1) Total Squared Difference: Let $\phi[g] = \int g^2 d\nu$, where $\phi : L^2(\nu) \to \mathbb{R}$, and let $g, f, a \in L^2(\nu)$. Then

$$\begin{aligned}\phi[g+a] - \phi[g] &= \int (g+a)^2 d\nu - \int g^2 d\nu \\ &= 2\int ga d\nu + \int a^2 d\nu.\end{aligned}$$

Because

$$\frac{\int a^2 d\nu}{\|a\|_{L^2(\nu)}} = \frac{\|a\|^2_{L^2(\nu)}}{\|a\|_{L^2(\nu)}} = \|a\|_{L^2(\nu)} \to 0$$

as $a \to 0$ in $L^2(\nu)$,

$$\delta\phi[g; a] = 2\int ga d\nu,$$

which is a continuous linear functional in a. To show that ϕ is strictly convex we show that ϕ is strongly positive. When the second variation $\delta^2\phi$ and the third variation $\delta^3\phi$ exist, they are described by

$$\begin{aligned}\triangle\phi[f; a] &= \delta\phi[f; a] + \frac{1}{2}\delta^2\phi[f; a, a] \\ &\quad + \epsilon[f, a]\,\|a\|^2_{L^p(\nu)} \qquad (3) \\ &= \delta\phi[f; a] + \frac{1}{2}\delta^2\phi[f; a, a] \\ &\quad + \frac{1}{6}\delta^3\phi[f; a, a, a] \\ &\quad + \epsilon[f, a]\,\|a\|^3_{L^p(\nu)},\end{aligned}$$

where $\epsilon[f, a] \to 0$ as $\|a\|_{L^p(\nu)} \to 0$. The quadratic functional $\delta^2\phi[f; a, a]$ defined on normed linear space $L^p(\nu)$ is ***strongly positive*** if there exists a constant $k > 0$ such that $\delta^2\phi[f; a, a] \geq k\,\|a\|^2_{L^p(\nu)}$ for all $a \in L^2(\nu)$. By definition of the second Fréchet derivative,

$$\begin{aligned}\delta^2\phi[g; b, a] &= \delta\phi[g+b; a] - \delta\phi[g; a] \\ &= 2\int (g+b)a d\nu - 2\int ga d\nu \\ &= 2\int ba d\nu.\end{aligned}$$

Thus $\delta^2\phi[g; b, a]$ is a quadratic form, where $\delta^2\phi$ is actually independent of g and strongly positive since

$$\delta^2\phi[g; a, a] = 2\int a^2 d\nu = 2\|a\|^2_{L^2(\nu)}$$

for all $a \in L^2(\nu)$, which implies that ϕ is strictly convex and

$$\begin{aligned}d_\phi[f, g] &= \int f^2 d\nu - \int g^2 d\nu - 2\int g(f-g) d\nu \\ &= \int (f-g)^2 d\nu \\ &= \|f-g\|^2_{L^2(\nu)}.\end{aligned}$$

2) Squared Bias: Under definition (2), squared bias is a Bregman divergence, this we have not previously seen noted in the literature despite the importance of minimizing bias in estimation [21].

Let $\phi[g] = \left(\int g d\nu\right)^2$, where $\phi : L^1(\nu) \to \mathbb{R}$. In this case

$$\begin{aligned}\phi[g+a] - \phi[g] &= \left(\int g d\nu + \int a d\nu\right)^2 - \left(\int g d\nu\right)^2 \\ &= 2\int g d\nu \int a d\nu + \left(\int a d\nu\right)^2. \qquad (4)\end{aligned}$$

Note that $2\int g d\nu \int a d\nu$ is a continuous linear functional on $L^1(\nu)$ and $\left(\int a d\nu\right)^2 \leq \|a\|^2_{L^1(\nu)}$, so that

$$0 \leq \frac{\left(\int a d\nu\right)^2}{\|a\|_{L^1(\nu)}} \leq \frac{\|a\|^2_{L^1(\nu)}}{\|a\|_{L^1(\nu)}} = \|a\|_{L^1(\nu)}.$$

Thus from (4) and the definition of the Fréchet derivative,

$$\delta\phi[g; a] = 2\int g d\nu \int a d\nu.$$

By the definition of the second Fréchet derivative,

$$\begin{aligned}\delta^2\phi[g; b, a] &= \delta\phi[g+b; a] - \delta\phi[g; a] \\ &= 2\int (g+b) d\nu \int a d\nu - 2\int g d\nu \int a d\nu \\ &= 2\int b d\nu \int a d\nu\end{aligned}$$

is another quadratic form, and $\delta^2\phi$ is independent of g.

Then $\delta^2\phi[g; a, a]$ is strongly positive because,

$$\delta^2\phi[g; a, a] = 2\left(\int a d\nu\right)^2 = 2\|a\|^2_{L^1(\nu)} \geq 0$$

for $a \in L^1(\nu)$, and thus ϕ is in fact strictly convex. The Bregman divergence is thus

$$\begin{aligned}&d_\phi[f, g] \\ &= \left(\int f d\nu\right)^2 - \left(\int g d\nu\right)^2 - 2\int g d\nu \int (f-g) d\nu \\ &= \left(\int f d\nu\right)^2 + \left(\int g d\nu\right)^2 - 2\int g d\nu \int f d\nu \\ &= \left(\int (f-g) d\nu\right)^2 \\ &\leq \|f-g\|^2_{L^1(\nu)}.\end{aligned}$$

3) Relative Entropy of Simple Functions: Denote by $\mathcal{S}$ the collection of all integrable simple functions on the measure space $(\mathbb{R}^d, \Omega, \nu)$, that is, the set of functions which can be written as a finite linear combination of indicator functions such that if $g \in \mathcal{S}$, g can be expressed,

$$g(x) = \sum_{i=0}^{t} \alpha_i I_{T_i}; \quad \alpha_0 = 0,$$

where I_{T_i} is the indicator function of the set T_i, $\{T_i\}_{i=1}^t$ is a collection of mutually disjoint sets of finite measure and $T_0 = \mathbb{R}^d \setminus \bigcup_{i=1}^t T_i$. We adopt the convention that T_0 is the set on which g is zero and therefore $\alpha_i \neq 0$ if $i \neq 0$.

Consider the normed vector space $(\mathcal{S}, \|\cdot\|_{L^\infty(\nu)})$ and let $\mathcal{W}$ be the subset (not necessarily a vector subspace) of non-negative functions in this normed space:

$$\mathcal{W} = \{g \in \mathcal{S} \text{ subject to } g \geq 0\}.$$

If $g \in \mathcal{W}$ then

$$\int_{\mathbb{R}^d} g \ln g d\nu = \sum_{i=1}^{t} \int_{T_i} \alpha_i \ln \alpha_i d\nu = \sum_{i=1}^{t} \alpha_i \ln \alpha_i \nu(T_i), \quad (5)$$

since $0 \ln 0 = 0$. Define the functional ϕ on $\mathcal{W}$,

$$\phi[g] = \int_{\mathbb{R}^d} g \ln g \, d\nu, \quad g \in \mathcal{W}. \quad (6)$$

The functional ϕ is not Fréchet-differentiable at g because in general it cannot be guaranteed that $g + h$ is non-negative on the set where $g = 0$ for all perturbing functions h in the underlying normed vector space $\left(\mathcal{S}, \|\cdot\|_{L^\infty(\nu)}\right)$ with norm smaller than any prescribed $\epsilon > 0$. However, a generalized Gâteaux derivative can be defined if we limit the perturbing function h to a vector subspace. To that end, let $\mathcal{G}$ be the subspace of $\left(\mathcal{S}, \|\cdot\|_{L^\infty(\nu)}\right)$ defined by

$$\mathcal{G} = \{f \in \mathcal{S} \text{ subject to } f \, d\nu \ll g \, d\nu\}.$$

It is straightforward to show that $\mathcal{G}$ is a vector space. We define the generalized Gâteaux derivative of ϕ at $g \in \mathcal{W}$ to be the linear operator $\delta_G \phi[g; \cdot]$ if

$$\lim_{\substack{\|h\|_{L^\infty(\nu)} \to 0 \\ h \in \mathcal{G}}} \frac{|\phi[g+h] - \phi[g] - \delta_G \phi[g; h]|}{\|h\|_{L^\infty(\nu)}} = 0. \quad (7)$$

Note, that $\delta_G \phi[g; \cdot]$ is not linear in general, but it is on the vector space $\mathcal{G}$. In general, if $\mathcal{G}$ is the entire underlying vector space then (7) is the Fréchet derivative, and if $\mathcal{G}$ is the span of only one element from the underlying vector space then (7) is the Gâteaux derivative. Here, we have generalized the Gâteaux derivative for the present case that $\mathcal{G}$ is a subspace of the underlying vector space.

It remains to be shown that given the functional (6), the derivative (7) exists and yields a Bregman divergence corresponding to the usual notion of relative entropy. Consider the possible solution

$$\delta_G \phi[g; h] = \int_{\mathbb{R}^d} (1 + \ln g) h d\nu, \quad (8)$$

which coupled with (6) does yield relative entropy. It remains to be shown only that (8) satisfies (7). Note that

$$\begin{aligned} \phi[g+h] - \phi[g] - \delta_G \phi[g; h] &= \int_{\mathbb{R}^d} (h+g) \ln \frac{h+g}{g} - h d\nu \\ &= \int_E (h+g) \ln \frac{h+g}{g} - h d\nu, \end{aligned} \quad (9)$$

where E is the set on which g is not zero.

Because $g \in \mathcal{W}$, there are $m, M > 0$ such that $m \leq g \leq M$ on E. Let $h \in \mathcal{G}$ be such that $\|h\|_{L^\infty(\nu)} \leq m$, then $g + h \geq 0$. Our goal is to show that the expression

$$\frac{\phi[g+h] - \phi[g] - \delta_G \phi[g; h]}{\|h\|_{L^\infty(\nu)}} \quad (10)$$

is non-negative and that it is bounded above by a bound that goes to 0 as $\|h\|_{L^\infty(\nu)} \to 0$. We start by bounding the integrand from above using the inequality $\ln x \leq x - 1$:

$$(h+g) \ln \frac{h+g}{g} - h \leq (h+g) \frac{h}{g} - h = \frac{h^2}{g}.$$

Then since $h^2/g \leq (\|h\|_{L^\infty(\nu)})^2 / m$,

$$\begin{aligned} \frac{\phi[g+h] - \phi[g] - \delta_G \phi[g; h]}{\|h\|_{L^\infty(\nu)}} &\leq \frac{1}{\|h\|_{L^\infty(\nu)}} \int_E \frac{h^2}{g} d\nu \\ &\leq \frac{\nu(E)}{m} \|h\|_{L^\infty(\nu)}. \end{aligned}$$

Since g is integrable $\nu(E) < \infty$ and the right hand side goes to 0 as $\|h\|_{L^\infty(\nu)} \to 0$.

Next, in order to show that (10) is non-negative we have to prove that the integral (9) is not negative. To do so, we normalize the measure and apply Jensen's inequality. Take the first term of the integrand of (9),

$$\begin{aligned} &\int_E (h+g) \ln \frac{h+g}{g} d\nu \\ &= \int_E \frac{h+g}{g} \left(\ln \frac{h+g}{g} \right) g d\nu, \\ &= \|g\|_{L^1(\nu)} \int_E \frac{h+g}{g} \ln \frac{h+g}{g} \frac{g}{\|g\|_{L^1(\nu)}} d\nu \\ &= \|g\|_{L^1(\nu)} \int_E \lambda \left(\frac{h+g}{g} \right) d\tilde{\nu}, \end{aligned}$$

where the normalized measure $d\tilde{\nu} = \frac{g}{\|g\|_{L^1(\nu)}} d\nu$ is a probability measure and $\lambda(x) = x \ln x$ is a convex function on $(0, \infty)$. By Jensen's inequality and then changing the measure back to

$d\nu$,

$$\begin{aligned}
\|g\|_{L^1(\nu)} \int_E \lambda\left(\frac{h+g}{g}\right) d\tilde{\nu} & \\
\geq & \|g\|_{L^1(\nu)} \lambda\left(\int_E \frac{h+g}{g} d\tilde{\nu}\right) \\
= & \|g\|_{L^1(\nu)} \lambda\left(\frac{\|g+h\|_{L^1(\nu)}}{\|g\|_{L^1(\nu)}}\right) \\
= & \|g+h\|_{L^1(\nu)} \ln\left(\frac{\|g+h\|_{L^1(\nu)}}{\|g\|_{L^1(\nu)}}\right) \\
\geq & \|g+h\|_{L^1(\nu)} \left(1 - \frac{\|g\|_{L^1(\nu)}}{\|g+h\|_{L^1(\nu)}}\right) \\
= & \|g+h\|_{L^1(\nu)} - \|g\|_{L^1(\nu)} = \int_E h d\nu,
\end{aligned}$$

where we used the fact that $\ln \frac{1}{x} \geq 1 - x$ for all $x > 0$. By combining these two latest results we find that

$$\int_E (h+g) \ln \frac{h+g}{g} d\nu \geq \int_E h d\nu,$$

so equivalently (9) is always non-negative. This fact also confirms that the resulting relative entropy $d_\phi[f, g]$ is always non-negative, because (9) is $d_\phi[f, g]$ if one sets $h = f - g$.

Lastly, one must show that the functional defined in (6) is strictly convex. Again we will show this by showing that the second variation of $\phi[g]$ is strongly positive. Let $f \in \mathcal{G}$ and $\|f\|_{L^\infty(\nu)} \leq m$. Using the Taylor expansion of $\ln$ one can express,

$$\begin{aligned}
\delta_G \phi[g+f; h] - \delta_G \phi[g; h] &= \int_E h \ln\left(1 + \frac{f}{g}\right) d\nu \\
&= \int_E h \frac{f}{g} d\nu + \epsilon[f, g] \|f\|_{L^\infty(\nu)},
\end{aligned}$$

where $\epsilon[f, g]$ goes to 0 as $\|f\|_{L^\infty(\nu)} \to 0$ because

$$\|\epsilon[f, g]\|_{L^\infty(\nu)} \leq \frac{\nu(E)\|h\|_{L^\infty(\nu)}}{2M^2} \|f\|_{L^\infty(\nu)}.$$

Therefore

$$\delta_G^2 \phi[g; h, f] = \int_E h \frac{f}{g} d\nu,$$

and

$$\begin{aligned}
\delta_G^2 \phi[g; h, h] &= \int_E \frac{h^2}{g} d\nu \\
&\geq \frac{1}{M} \|h\|_{L^1(\nu)}.
\end{aligned}$$

Thus $\delta_G^2 \phi[g; h, h]$ is strongly positive.

B. Relationship to Other Bregman Divergence Definitions

Two propositions establish the relationship of the functional Bregman divergence to other Bregman divergence definitions.

Proposition I.2 (Functional Bregman Divergence Generalizes Vector Bregman Divergence)**.** *The functional definition (2) is a generalization of the standard vector Bregman divergence*

$$d_{\tilde{\phi}}(x, y) = \tilde{\phi}(x) - \tilde{\phi}(y) - \nabla \tilde{\phi}(y)^T (x - y), \qquad (11)$$

where $x, y \in \mathbb{R}^n$, and $\tilde{\phi} : \mathbb{R}^n \to \mathbb{R}$ is strictly convex and twice differentiable.

Jones and Byrne describe a general class of divergences between functions using a pointwise formulation [7]. Csiszár specialized the pointwise formulation to a class of divergences he termed *Bregman distances* $B_{s,\nu}$ [17], where given a σ-finite measure space (X, Ω, ν), and non-negative measurable functions $f(x)$ and $g(x)$, $B_{s,\nu}(f, g)$ equals

$$\int s(f(x)) - s(g(x)) - s'(g(x))(f(x) - g(x)) d\nu(x). \quad (12)$$

The function $s : (0, \infty) \to \mathbb{R}$ is constrained to be differentiable and strictly convex, and the limit $\lim_{x \to 0} s(x)$ and $\lim_{x \to 0} s'(x)$ must exist, but not necessarily be finite. The function s plays a role similar to the function ϕ in the functional Bregman divergence; however, s acts on the range of the functions f, g, whereas ϕ acts on the functions f, g.

Proposition I.3 (Functional Definition Generalizes Pointwise Definition)**.** *Given a pointwise Bregman divergence as per (12), an equivalent functional Bregman divergence can be defined as per (2) if the measure ν is finite. However, given a functional Bregman divergence $d_\phi[f, g]$, there is not necessarily an equivalent pointwise Bregman divergence.*

C. Properties of the Functional Bregman Divergence

The Bregman divergence for random variables has some well-known properties, as reviewed in [11, Appendix A]. Here, we note that the same properties hold for the functional Bregman divergence (2). We give complete proofs in [18].

1. Non-negativity: The functional Bregman divergence is non-negative: $d_\phi[f, g] \geq 0$ for all admissible inputs.

2. Convexity: The Bregman divergence $d_\phi[f, g]$ is always convex with respect to f.

3. Linearity: The functional Bregman divergence is linear such that,

$$d_{(c_1\phi_1 + c_2\phi_2)}[f, g] = c_1 d_{\phi_1}[f, g] + c_2 d_{\phi_2}[f, g].$$

4. Equivalence Classes: Partition the set of strictly convex, differentiable functionals $\{\phi\}$ into classes such that ϕ_1 and ϕ_2 belong to the same class if $d_{\phi_1}[f, g] = d_{\phi_2}[f, g]$ for all $f, g \in \mathcal{A}$. For brevity we will denote $d_{\phi_1}[f, g]$ simply by d_{ϕ_1}. Let $\phi_1 \sim \phi_2$ denote that ϕ_1 and ϕ_2 belong to the same class, then $\sim$ is an equivalence relation in that is satisfies the properties of *reflexivity* ($d_{\phi_1} = d_{\phi_1}$), *symmetry* (if $d_{\phi_1} = d_{\phi_2}$, then $d_{\phi_2} = d_{\phi_1}$), and *transitivity* (if $d_{\phi_1} = d_{\phi_2}$ and $d_{\phi_2} = d_{\phi_3}$, then $d_{\phi_1} = d_{\phi_3}$). Further, if $\phi_1 \sim \phi_2$, then they differ only by an affine transformation.

5. Linear Separation: The locus of admissible functions $f \in L^p(\nu)$ that are equidistant from two fixed functions $g_1, g_2 \in L^p(\nu)$ in terms of functional Bregman divergence form a hyperplane.

6. Dual Divergence: Given a pair (g, ϕ) where $g \in L^p(\nu)$ and ϕ is a strictly convex twice-continuously Fréchet-differentiable functional, then the function-functional pair (G, ψ) is the Legendre transform of (g, ϕ) [19], if

$$\phi[g] = -\psi[G] + \int g(x)G(x)d\nu(x), \quad (13)$$

$$\delta\phi[g; a] = \int G(x)a(x)d\nu(x), \quad (14)$$

where ψ is a strictly convex twice-continuously Fréchet-differentiable functional, and $G \in L^q(\nu)$, where $\frac{1}{p} + \frac{1}{q} = 1$.

Given Legendre transformation pairs $f, g \in L^p(\nu)$ and $F, G \in L^q(\nu)$,

$$d_\phi[f, g] = d_\psi[G, F].$$

7. Generalized Pythagorean Inequality: For any admissible $f, g, h \in L^p(\nu)$,

$$d_\phi[f, h] = d_\phi[f, g] + d_\phi[g, h] + \delta\phi[g; f - g] - \delta\phi[h; f - g].$$

II. Minimum Expected Bregman Divergence

Consider two sets of functions (or distributions), $\mathcal{M}$ and $\mathcal{A}$. Let $F \in \mathcal{M}$ be a random function with realization f. Suppose there exists a probability distribution P_F over the set $\mathcal{M}$, such that $P_F(f)$ is the probability of $f \in \mathcal{M}$. For example, consider the set of Gaussian distributions, and given samples drawn independently and identically from a randomly selected Gaussian distribution N, the data imply a posterior probability $P_N(\mathcal{N})$ for each possible generating realization of a Gaussian distribution $\mathcal{N}$. The goal is to find the function $g^* \in \mathcal{A}$ that minimizes the expected Bregman divergence between the random function F and any function $g \in \mathcal{A}$. The following theorem shows that if the set of possible minimizers $\mathcal{A}$ includes $E_{P_F}[F]$, then $g^* = E_{P_F}[F]$ minimizes the expectation of any Bregman divergence. Note the theorem requires slightly stronger conditions on ϕ than the definition of the Bregman divergence (2) requires.

Theorem II.1 (Minimizer of the Expected Bregman Divergence)**.** *Let $\delta^2\phi[f; a, a]$ be strongly positive and let $\phi \in \mathcal{C}^3(L^1(\nu); \mathbb{R})$ be a three-times continuously Fréchet-differentiable functional on $L^1(\nu)$. Let $\mathcal{M}$ be a set of functions that lie on a manifold M, and have associated measure dM such that integration is well-defined. Suppose there is a probability distribution P_F defined over the set $\mathcal{M}$. Let $\mathcal{A}$ be a set of functions that includes $E_{P_F}[F]$ if it exists. Suppose the function g^* minimizes the expected Bregman divergence between the random function F and any function $g \in \mathcal{A}$ such that*

$$g^* = \arg\inf_{g \in \mathcal{A}} E_{P_F}[d_\phi(F, g)].$$

Then, if g^* exists, it is given by

$$g^* = E_{P_F}[F]. \quad (15)$$

A. Bayesian Estimation

Theorem II.1 can be applied to a set of distributions to find the Bayesian estimate of a distribution given a posterior or likelihood. For parametric distributions parameterized by $\theta \in \mathbb{R}^n$, a probability measure $\Lambda(\theta)$, and some risk function $R(\theta, \psi)$, $\psi \in \mathbb{R}^n$, the Bayes estimator is defined [22] as

$$\hat{\theta} = \arg\inf_{\psi \in \mathbb{R}^n} \int R(\theta, \psi)d\Lambda(\theta). \quad (16)$$

That is, the Bayes estimator minimizes some expected risk in terms of the parameters. It follows from recent results [16] that $\hat{\theta} = E[\Theta]$ if the risk R is a Bregman divergence, where Θ is the random variable whose realization is θ; this property has been previously noted [8], [10].

The principle of Bayesian estimation can be applied to the distributions themselves rather than to the parameters:

$$\hat{g} = \arg\inf_{g \in \mathcal{A}} \int_M R(f, g)P_F(f)dM, \quad (17)$$

where $P_F(f)$ is a probability measure on the distributions $f \in \mathcal{M}$, dM is a measure for the manifold M, and $\mathcal{A}$ is either the space of all distributions or a subset of the space of all distributions, such as the set $\mathcal{M}$. When the set $\mathcal{A}$ includes the distribution $E_{P_F}[F]$ and the risk function R in (17) is a functional Bregman divergence, then Theorem II.1 establishes that $\hat{g} = E_{P_F}[F]$.

For example, in recent work, two of the authors derived the mean class posterior distribution for each class for a Bayesian quadratic discriminant analysis classifier, and showed that the classification results were superior to parameter-based Bayesian quadratic discriminant analysis [23].

Of particular interest for estimation problems are the Bregman divergence examples given in Section I-A: total squared difference (mean squared error) is a popular risk function in regression [21]; minimizing relative entropy leads to useful theorems for large deviations and other statistical subfields [24]; and analyzing bias is a common approach to characterizing and understanding statistical learning algorithms [21].

B. Case Study: Estimating a Scaled Uniform Distribution

As an illustration of the theorem, we present and compare different estimates of a scaled uniform distribution given independent and identically drawn samples. Let the set of uniform distributions over $[0, \theta]$ for $\theta \in \mathbb{R}^+$ be denoted by $\mathcal{U}$. Given independent and identically distributed samples $X_1, X_2, \ldots, X_n$ drawn from an unknown uniform distribution $f \in \mathcal{U}$, the generating distribution is to be estimated. The risk function R is taken to be squared error or total squared error depending on context.

1) Bayesian Parameter Estimate: Depending on the choice of the probability measure $\Lambda(\theta)$, the integral (16) may not be finite; for example, using the likelihood of θ with Lebesgue measure the integral is not finite. A standard solution is to use a gamma prior on θ and Lebesgue measure. Let Θ be a random parameter with realization θ, let the gamma distribution have parameters t_1 and t_2, and denote the maximum of the data as

$X_{\max} = \max\{X_1, X_2, \ldots, X_n\}$. Then a Bayesian estimate is formulated [22, p. 240, 285]:

$$\begin{aligned} E[\Theta|\{X_1, X_2, \ldots, X_n\}, t_1, t_2] \\ = \frac{\int_{X_{\max}}^{\infty} \theta \frac{1}{\theta^{n+t_1+1}} e^{\frac{-1}{\theta t_2}} d\theta}{\int_{X_{\max}}^{\infty} \frac{1}{\theta^{n+t_1+1}} e^{\frac{-1}{\theta t_2}} d\theta}. \end{aligned} \quad (18)$$

The integrals can be expressed in terms of the chi-squared random variable I_v^2 with v degrees of freedom:

$$\begin{aligned} E[\Theta|\{X_1, X_2, \ldots, X_n\}, t_1, t_2] = \\ \frac{1}{t_2(n+t_1-a)} \frac{P(\chi^2_{2(n+t_1-1)} < \frac{2}{t_2 X_{\max}})}{P(\chi^2_{2(n+t_1)} < \frac{2}{t_2 X_{\max}})}. \end{aligned} \quad (19)$$

Note that (16) presupposes that the best solution is also a uniform distribution.

2) Bayesian Uniform Distribution Estimate: If one restricts the minimizer of (17) to be a uniform distribution, then (17) is solved with $\mathcal{A} = \mathcal{U}$. Because the set of uniform distributions does not generally include its mean, Theorem II.1 does not apply, and thus different Bregman divergences may give different minimizers for (17). Let P_F be the likelihood of the data (no prior is assumed over the set $\mathcal{U}$), and use the Fisher information metric ([25]–[27]) for dM. Then the solution to (17) is the uniform distribution on $[0, 2^{1/n} X_{\max}]$. Using Lebesgue measure instead gives a similar result: $[0, 2^{1/(n+1/2)} X_{\max}]$. We were unable to find these estimates in the literature, and so their derivations are presented in the appendix.

3) Unrestricted Bayesian Distribution Estimate: When the only restriction placed on the minimizer g in (17) is that g be a distribution, then one can apply Theorem II.1 and solve directly for the expected distribution $E_{P_F}[F]$. Let P_F be the likelihood of the data (no prior is assumed over the set $\mathcal{U}$), and use the Fisher information metric for dM. Solving (15), noting that the uniform probability of x is $f(x) = 1/a$ if $x \leq a$ and zero otherwise, and the likelihood of the n drawn points is $(1/X_{\max})^n$ if $a \geq X_{\max}$ and zero otherwise,

$$\begin{aligned} g^*(x) &= \frac{\int_{\max(x, X_{\max})}^{\infty} \left(\frac{1}{a}\right) \left(\frac{1}{a^n}\right) \left(\frac{da}{a}\right)}{\int_{X_{\max}}^{\infty} \frac{1}{a^n} \frac{da}{a}} \\ &= \frac{n \left(X_{\max}\right)^n}{(n+1)[\max(x, X_{\max})]^{n+1}}. \end{aligned} \quad (20)$$

III. Further Discussion and Open Questions

We have defined a general Bregman divergence for functions and distributions that can provide a foundation for results in statistics, information theory and signal processing. Theorem II.1 is important for these fields because it ties Bregman divergences to expectation. As shown in Section II-A, Theorem II.1 can be directly applied to distributions to show that Bayesian distribution estimation simplifies to expectation when the risk function is a Bregman divergence and the minimizing distribution is unrestricted.

It is common in Bayesian estimation to interpret the prior as representing some actual prior knowledge, but in fact prior knowledge often is not available or is difficult to quantify. Another approach is to use a prior to capture coarse information from the data that may be used to stabilize the estimation [9], [23]. In practice, priors are sometimes chosen in Bayesian estimation to tame the tail of likelihood distributions so that expectations will exist when they might otherwise be infinite [22]. This mathematically convenient use of priors adds estimation bias that may be unwarranted by prior knowledge. An alternative to mathematically convenient priors is to formulate the estimation problem as a minimization of an expected Bregman divergence between the unknown distribution and the estimated distribution, and restrict the set of distributions that can be the minimizer to be a set for which the Bayesian integral exist. Open questions are how such restrictions trade-off bias for reduced variance, and how to find or define an "optimal" restricted set of distributions for this estimation approach.

Finally, there are some results for the standard vector Bregman divergence that have not been extended here. It has been shown that a standard vector Bregman divergence must be the risk function in order for the mean to be the minimizer of an expected risk [16, Theorems 3 and 4]. The proof of that result relies heavily on the discrete nature of the underlying vectors, and it remains an open question as to whether a similar result holds for the functional Bregman divergence. Another result that has been shown for the vector case but remains an open question in the functional case is convergence in probability [16, Theorem 2].

Acknowledgments

This work was funded in part by the Office of Naval Research, Code 321, Grant # N00014-05-1-0843. The authors thank Inderjit Dhillon, Castedo Ellerman, and Galen Shorack for helpful discussions.

Appendix: Proofs

A. Proof of Proposition I.2

We give a constructive proof that there is a corresponding functional Bregman divergence $d_\phi[f, g]$ for a specific choice of $\phi : L^1(\nu) \to \mathbb{R}$, where $\nu = \sum_{i=1}^n \delta_{c_i}$ and $f, g \in L^1(\nu)$. Here, δ_x denotes the Dirac measure such that all mass is concentrated at x, and $\{c_1, c_2, \ldots, c_n\}$ is a collection of n distinct points in $\mathbb{R}^d$.

For any $x \in \mathbb{R}^n$, define $\phi[f] = \tilde{\phi}(x_1, x_2, \ldots, x_n)$, where $f(c_1) = x_1, f(c_2) = x_2, \ldots, f(c_n) = x_n$. Then the difference is

$$\begin{aligned} \Delta\phi[f; a] &= \phi[f + a] - \phi[f] \\ &= \tilde{\phi}\left((f+a)(c_1), \ldots, (f+a)(c_n)\right) - \tilde{\phi}\left(x_1, \ldots, x_n\right) \\ &= \tilde{\phi}\left(x_1 + a(c_1), \ldots, x_n + a(c_n)\right) - \tilde{\phi}\left(x_1, \ldots, x_n\right). \end{aligned}$$

Let a_i be short hand for $a(c_i)$, and use the Taylor expansion for functions of several variables to yield

$$\Delta\phi[f; a] = \nabla\tilde{\phi}(x_1, \ldots, x_n)^T (a_1, \ldots, a_n) + \epsilon[f, a]\|a\|_{L^1}.$$

Therefore,

$$\delta\phi[f; a] = \nabla\tilde{\phi}(x_1, \ldots, x_n)^T (a_1, \ldots, a_n) = \nabla\tilde{\phi}(x)^T a,$$

where $x = (x_1, x_2, \ldots, x_n)$ and $a = (a_1, \ldots, a_n)$. Thus, from (3), the functional Bregman divergence definition (2) for ϕ is equivalent to the standard vector Bregman divergence:

$$\begin{aligned} d_{\tilde{\phi}}[f, g] &= \phi[f] - \phi[g] - \delta\phi[g; f - g] \\ &= \tilde{\phi}(x) - \tilde{\phi}(y) - \nabla\tilde{\phi}(y)^T(x - y). \end{aligned} \quad (21)$$

B. Proof of Proposition I.3

First, we give a constructive proof of the first part of the proposition by showing that given a $B_{s,\nu}$, there is an equivalent functional divergence d_ϕ. Then, the second part of the proposition is shown by example: we prove that the squared bias functional Bregman divergence given in Section I-A.2 is a functional Bregman divergence that cannot be defined as a pointwise Bregman divergence.

Note that the integral to calculate $B_{s,\nu}$ is not always finite. To ensure finite $B_{s,\nu}$, we explicitly constrain $\lim_{x\to 0} s'(x)$ and $\lim_{x\to 0} s(x)$ to be finite. From the assumption that s is strictly convex, s must be continuous on $(0, \infty)$. Recall from the assumptions that the measure ν is finite, and that the function s is differentiable on $(0, \infty)$.

Given a $B_{s,\nu}$, define the continuously differentiable function

$$\tilde{s}(x) = \begin{cases} s(x) & x \geq 0 \\ -s(-x) + 2s(0) & x < 0. \end{cases}$$

Specify $\phi : L^\infty(\nu) \to \mathbb{R}$ as

$$\phi[f] = \int_X \tilde{s}(f(x))d\nu.$$

Note that if $f \geq 0$,

$$\phi[f] = \int_X s(f(x))d\nu.$$

Because $\tilde{s}$ is continuous on $\mathbb{R}$, $\tilde{s}(f) \in L^\infty(\nu)$ whenever $f \in L^\infty(\nu)$, so the above integrals always make sense.

It remains to be shown that $\delta\phi[f; \cdot]$ completes the equivalence when $f \geq 0$. For $h \in L^\infty(\nu)$,

$$\begin{aligned} \phi[f + h] - \phi[f] &= \int_X \tilde{s}(f(x) + h(x))d\nu - \int_X s(f(x))d\nu \\ &= \int_X \tilde{s}(f(x) + h(x)) - s(f(x))d\nu \\ &= \int_X \tilde{s}'(f(x))h(x) + \epsilon\left(f(x), h(x)\right)h(x)d\nu \\ &= \int_X s'(f(x))h(x) + \epsilon\left(f(x), h(x)\right)h(x)d\nu, \end{aligned}$$

where we used the fact that

$$\begin{aligned} &\tilde{s}(f(x) + h(x)) \\ &\quad = \tilde{s}(f(x)) + (\tilde{s}'(f(x)) + \epsilon(f(x), h(x)))\,h(x) \\ &\quad = s(f(x)) + (s'(f(x)) + \epsilon(f(x), h(x)))\,h(x), \end{aligned}$$

because $f \geq 0$. On the other hand, if $h(x) = 0$ then $\epsilon(f(x), h(x)) = 0$, and if $h(x) \neq 0$ then

$$|\epsilon(f(x), h(x))| \leq \left|\frac{\tilde{s}(f(x) + h(x)) - \tilde{s}(f(x))}{h(x)}\right| + |s'(f(x))|.$$

Suppose $\{h_n\} \subset L^\infty(\nu)$ such that $h_n \to 0$. Then there is a measurable set E such that its complement is of measure 0 and $h_n \to 0$ uniformly on E. There is some $N > 0$ such that for any $n > N$, $|h_n(x)| \leq \epsilon$ for all $x \in E$. Without loss of generality, assume that there is some $M > 0$ such that for all $x \in E$, $|f(x)| \leq M$. Since $\tilde{s}$ is continuously differentiable, there is a $K > 0$ such that $\max\{\tilde{s}'(t) \text{ subject to } t \in [-M - \epsilon, M + \epsilon]\} \leq K$, and by the mean value theorem

$$\left|\frac{\tilde{s}(f(x) + h(x)) - \tilde{s}(f(x))}{h(x)}\right| \leq K,$$

for almost all $x \in X$. Then

$$|\epsilon(f(x), h(x))| \leq 2K,$$

except on a set of measure 0. The fact that $h(x) \to 0$ almost everywhere implies that $|\epsilon(f(x), h(x))| \to 0$ almost everywhere, and by Lebesgue's dominated convergence theorem, the corresponding integral goes to 0. As a result, the Fréchet derivative of ϕ is

$$\delta\phi[f; h] = \int_X s'(f(x))h(x)d\nu. \quad (22)$$

Thus the functional Bregman divergence is equivalent to the given pointwise $B_{s,\nu}$.

We additionally note that the assumptions that $f \in L^\infty(\nu)$ and that the measure ν is finite are necessary for this proof. Counterexamples can be constructed if $f \in L^p$ or $\nu(X) = \infty$ such that the Fréchet derivative of ϕ does not obey (22). This concludes the first part of the proof.

To show that the squared bias functional Bregman divergence given in Section I-A.2 is an example of a functional Bregman divergence that cannot be defined as a pointwise Bregman divergence we prove that the converse statement leads to a contradiction.

Suppose (X, Σ, ν) and (X, Σ, μ) are measure spaces where ν is a non-zero σ-finite measure and that there is a differentiable function $f : (0, \infty) \to \mathbb{R}$ such that

$$\left(\int \xi d\nu\right)^2 = \int f(\xi)d\mu, \quad (23)$$

where $\xi \in L^1(\nu)$. Let $f(0) = \lim_{x\to 0} f(x)$, which can be finite or infinite, and let α be any real number. Then

$$\begin{aligned} \int f(\alpha\xi)d\mu = \left(\int \alpha\xi d\nu\right)^2 &= \alpha^2\left(\int \xi d\nu\right)^2 \\ &= \alpha^2 \int f(\xi)d\mu. \end{aligned}$$

Because ν is σ-finite, there is a measurable set E such that $0 < |\nu(E)| < \infty$. Let $X\backslash E$ denote the complement of E in X. Then

$$\begin{aligned} \alpha^2\nu^2(E) = \alpha^2\left(\int I_E d\nu\right)^2 &= \alpha^2 \int f(I_E)d\mu \\ &= \alpha^2 \int_{X\backslash E} f(0)d\mu + \alpha^2 \int_E f(1)d\mu \\ &= \alpha^2 f(0)\mu(X\backslash E) + \alpha^2 f(1)\mu(E). \end{aligned}$$

Also,

$$\alpha^2\nu^2(E) = \left(\int \alpha I_E d\nu\right)^2.$$

However,

$$\begin{aligned}\left(\int \alpha I_E d\nu\right)^2 &= \int f(\alpha I_E)d\mu \\ &= \int_{X\backslash E} f(\alpha I_E)d\mu + \int_E f(\alpha I_E)d\mu \\ &= f(0)\mu(X\backslash E) + f(\alpha)\mu(E);\end{aligned}$$

so one can conclude that

$$\begin{aligned}&\alpha^2 f(0)\mu(X\backslash E) + \alpha^2 f(1)\mu(E) \\ &= \quad f(0)\mu(X\backslash E) + f(\alpha)\mu(E). \end{aligned} \tag{24}$$

Apply equation (23) for $\xi = 0$ to yield

$$0 = \left(\int 0 d\nu\right)^2 = \int f(0)d\mu = f(0)\mu(X).$$

Since $|\nu(E)| > 0$, $\mu(X) \neq 0$, so it must be that $f(0) = 0$, and (24) becomes

$$\alpha^2\nu^2(E) = \alpha^2 f(1)\mu(E) = f(\alpha)\mu(E) \quad \forall\alpha \in \mathbb{R}.$$

The first equation implies that $\mu(E) \neq 0$. The second equation determines the function f completely:

$$f(\alpha) = f(1)\alpha^2.$$

Then (23) becomes

$$\left(\int \xi d\nu\right)^2 = \int f(1)\xi^2 d\mu.$$

Consider any two disjoint measurable sets, E_1 and E_2, with finite nonzero measure. Define $\xi_1 = I_{E_1}$ and $\xi_2 = I_{E_2}$. Then $\xi = \xi_1 + \xi_2$ and $\xi_1\xi_2 = I_{E_1}I_{E_2} = 0$. Equation (23) becomes

$$\int \xi_1 d\nu \int \xi_2 d\nu = f(1)\int \xi_1\xi_2 d\mu. \tag{25}$$

This implies the following contradiction:

$$\int \xi_1 d\nu \int \xi_2 d\nu = \nu(E_1)\nu(E_2) \neq 0, \tag{26}$$

but

$$f(1)\int \xi_1\xi_2 d\mu = 0. \tag{27}$$

C. Proof of Theorem II.1

Recall that for a functional J to have an extremum (minimum) at $f = \hat{f}$, it is necessary that

$$\delta J[f; a] = 0 \text{ and } \delta^2 J[f; a, a] \geq 0,$$

for $f = \hat{f}$ and for all admissible functions $a \in \mathcal{A}$. A sufficient condition for a functional $J[f]$ to have a minimum for $f = \hat{f}$ is that the first variation $\delta J[f; a]$ must vanish for $f = \hat{f}$, and its second variation $\delta^2 J[f; a, a]$ must be strongly positive for $f = \hat{f}$.

Let

$$\begin{aligned}J[g] &= E_{P_F}[d_\phi(F, g)] = \int_M d_\phi[f, g]P_F(f)dM \\ &= \int_M (\phi[f] - \phi[g] - \delta\phi[g; f - g])P_F(f)dM, \end{aligned} \tag{28}$$

where (28) follows by substituting the definition of Bregman divergence (2). Consider the increment

$$\begin{aligned}\Delta J[g; a] &= J[g + a] - J[g] \qquad (29)\\ &= -\int_M (\phi[g + a] - \phi[g])\, P_F(f)dM \\ &\quad -\int_M (\delta\phi[g + a; f - g - a] \\ &\quad -\delta\phi[g; f - g])\, P_F(f)dM, \qquad (30)\end{aligned}$$

where (30) follows from substituting (28) into (29). Using the definition of the differential of a functional given in (1), the first integrand in (30) can be written as

$$\phi[g + a] - \phi[g] \quad = \quad \delta\phi[g; a] + \epsilon[g, a]\,\|a\|_{L^1(\nu)}. \tag{31}$$

Take the second integrand of (30), and subtract and add $\delta\phi[g; f - g - a]$,

$$\begin{aligned}&\delta\phi[g + a; f - g - a] - \delta\phi[g; f - g] \\ &= \quad \delta\phi[g + a; f - g - a] - \delta\phi[g; f - g - a] \\ &\quad + \delta\phi[g; f - g - a] - \delta\phi[g; f - g] \\ &\overset{(a)}{=} \quad \delta^2\phi[g; f - g - a, a] + \epsilon[g, a]\,\|a\|_{L^1(\nu)} + \delta\phi[g; f - g] \\ &\quad - \delta\phi[g; a] - \delta\phi[g; f - g] \\ &\overset{(b)}{=} \quad \delta^2\phi[g; f - g, a] - \delta^2\phi[g; a, a] + \epsilon[g, a]\,\|a\|_{L^1(\nu)} \\ &\quad - \delta\phi[g; a] \end{aligned} \tag{32}$$

where (a) follows from the linearity of the third term, and (b) follows from the linearity of the first term. Substitute (31) and (32) into (30),

$$\begin{aligned}\triangle J[g; a] &= -\int_M \Big(\delta^2\phi[g; f - g, a] - \delta^2\phi[g; a, a] \\ &\quad + \epsilon[g, a]\,\|a\|_{L^1(\nu)}\Big)P_F(f)dM.\end{aligned}$$

Note that the term $\delta^2\phi[g; a, a]$ is of order $\|a\|^2_{L^1(\nu)}$, that is, $\left\|\delta^2\phi[g; a, a]\right\|_{L^1(\nu)} \leq m\,\|a\|^2_{L^1(\nu)}$ for some constant m. Therefore,

$$\lim_{\|a\|_{L^1(\nu)}\to 0} \frac{\|J[g + a] - J[g] - \delta J[g; a]\|_{L^1(\nu)}}{\|a\|_{L^1(\nu)}} = 0,$$

where,

$$\delta J[g; a] \quad = \quad -\int_M \delta^2\phi[g; f - g, a]P_F(f)dM. \tag{33}$$

For fixed a, $\delta^2\phi[g; \cdot, a]$ is a bounded linear functional in the second argument, so the integration and the functional can be interchanged in (33), which becomes

$$\delta J[g; a] = -\delta^2\phi\left[g; \int_M (f - g)\, P_F(f)dM, a\right].$$

Using the functional optimality conditions, $J[g]$ has an extremum for $g = \hat{g}$ if

$$\delta^2\phi\left[\hat{g}; \int_M (f - \hat{g})\, P_F(f) dM, a\right] = 0. \tag{34}$$

Set $a = \int_M (f - \hat{g})\, P(f) dM$ in (34) and use the assumption that the quadratic functional $\delta^2\phi[g; a, a]$ is strongly positive, which implies that the above functional can be zero if and only if $a = 0$, that is,

$$\begin{aligned} 0 &= \int_M (f - \hat{g}) P_F(f) dM, &(35)\\ \hat{g} &= E_{P_F}[F], &(36)\end{aligned}$$

where the last line holds if the expectation exists (i.e. if the measure is well-defined and the expectation is finite). Because a Bregman divergence is not necessarily convex in its second argument, it is not yet established that the above unique extremum is a minimum. To see that (36) is in fact a minimum of $J[g]$, from the functional optimality conditions it is enough to show that $\delta^2 J[\hat{g}; a, a]$ is strongly positive. To show this, for $b \in L^1(\nu)$, consider

$$\begin{aligned} &\delta J[g + b; a] - \delta J[g; a]\\ &\overset{(c)}{=} -\int_M \big(\delta^2\phi[g + b; f - g - b, a]\\ &\quad - \delta^2\phi[g; f - g, a]\big) P_F(f) dM\\ &\overset{(d)}{=} -\int_M \big(\delta^2\phi[g + b; f - g - b, a] - \delta^2\phi[g; f - g - b, a]\\ &\quad + \delta^2\phi[g; f - g - b, a] - \delta^2\phi[g; f - g, a]\big) P_F(f) dM\\ &\overset{(e)}{=} -\int_M \big(\delta^3\phi[g; f - g - b, a, b] + \epsilon[g, a, b]\, \|b\|_{L^1(\nu)}\\ &\quad + \delta^2\phi[g; f - g, a] - \delta^2\phi[g; b, a]\\ &\quad - \delta^2\phi[g; f - g, a]\big) P_F(f) dM\\ &\overset{(f)}{=} -\int_M \big(\delta^3\phi[g; f - g, a, b] - \delta^3\phi[g; b, a, b]\\ &\quad + \epsilon[g, a, b]\, \|b\|_{L^1(\nu)} - \delta^2\phi[g; b, a]\big) P_F(f) dM, \end{aligned} \tag{37}$$

where (c) follows from using integral (33); (d) from subtracting and adding $\delta^2\phi[g; f - g - b, a]$; (e) from the fact that the variation of the second variation of ϕ is the third variation of ϕ [28]; and (f) from the linearity of the first term and cancellation of the third and fifth terms. Note that in (37) for fixed a, the term $\delta^3\phi[g; b, a, b]$ is of order $\|b\|^2_{L^1(\nu)}$, while the first and the last terms are of order $\|b\|_{L^1(\nu)}$. Therefore,

$$\lim_{\|b\|_{L^1(\nu)} \to 0} \frac{\left\|\delta J[g + b; a] - \delta J[g; a] - \delta^2 J[g; a, b]\right\|_{L^1(\nu)}}{\|b\|_{L^1(\nu)}} = 0,$$

where

$$\begin{aligned} \delta^2 J[g; a, b] &= -\int_M \delta^3\phi[g; f - g, a, b] P_F(f) dM\\ &+ \int_M \delta^2\phi[g; a, b] P_F(f) dM. \end{aligned} \tag{38}$$

Substitute $b = a$, $g = \hat{g}$ and interchange integration and the continuous functional $\delta^3\phi$ in the first integral of (38), then

$$\begin{aligned} \delta^2 J[\hat{g}; a, a] &= -\delta^3\phi\left[\hat{g}; \int_M (f - \hat{g}) P_F(f) dM, a, a\right]\\ &\quad + \int_M \delta^2\phi[\hat{g}; a, a] P_F(f) dM\\ &= \int_M \delta^2\phi[\hat{g}; a, a] P_F(f) dM &(39)\\ &\geq \int_M k\, \|a\|^2_{L^1(\nu)}\, P_F(f) dM\\ &= k\, \|a\|^2_{L^1(\nu)} > 0, &(40)\end{aligned}$$

where (39) follows from (35), and (40) follows from the strong positivity of $\delta^2\phi[\hat{g}; a, a]$. Therefore, from (40) and the functional optimality conditions, $\hat{g}$ is the minimum.

D. Derivation of the Bayesian Distribution-based Uniform Estimate Restricted to a Uniform Minimizer

Let $f(x) = 1/a$ for all $0 \leq x \leq a$ and $g(x) = 1/b$ for all $0 \leq x \leq b$. Assume at first that $b > a$; then the total squared difference between f and g is

$$\begin{aligned} \int_x (f(x) - g(x))^2 dx &= a\left(\frac{1}{a} - \frac{1}{b}\right)^2 + (b - a)\left(\frac{1}{b}\right)^2\\ &= \frac{b - a}{ab}\\ &= \frac{|b - a|}{ab}, \end{aligned}$$

where the last line does not require the assumption that $b > a$.

In this case, the integral (17) is over the one-dimensional manifold of uniform distributions $\mathcal{U}$; a Riemannian metric can be formed by using the differential arc element to convert Lebesgue measure on the set $\mathcal{U}$ to a measure on the set of parameters a such that (17) is re-formulated in terms of the parameters for ease of calculation:

$$b^* = \arg\min_{b \in \mathbb{R}^+} \int_{a = X_{\max}}^{\infty} \frac{|b - a|}{ab} \frac{1}{a^n} \left\|\frac{df}{da}\right\|_2 da, \tag{41}$$

where $1/a^n$ is the likelihood of the n data points being drawn from a uniform distribution $[0, a]$, and the estimated distribution is uniform on $[0, b^*]$. The differential arc element $\left\|\frac{df}{da}\right\|_2$ can be calculated by expanding df/da in terms of the Haar orthonormal basis $\{\frac{1}{\sqrt{a}}, \phi_{jk}(x)\}$, which forms a complete orthonormal basis for the interval $0 \leq x \leq a$, and then the required norm is equivalent to the norm of the basis coefficients of the orthonormal expansion:

$$\left\|\frac{df}{da}\right\|_2 = \frac{1}{a^{3/2}}. \tag{42}$$

For estimation problems, the measure determined by the Fisher information metric may be more appropriate than Lebesgue measure [25]–[27]. Then

$$dM = |I(a)|^{\frac{1}{2}} da, \tag{43}$$

where I is the Fisher information matrix. For the one-dimensional manifold M formed by the set of scaled uniform

distributions $\mathcal{U}$, the Fisher information matrix is

$$\begin{aligned} I(a) &= E_X\left[-\left(\frac{d^2 \log \frac{1}{a}}{da^2}\right)\right] \\ &= \int_0^a \frac{1}{a^2}\frac{1}{a} dx = \frac{1}{a^2}, \end{aligned}$$

so that the differential element is $dM = \frac{da}{a}$.

We solve (17) using the Lebesgue measure (42); the solution with the Fisher differential element follows the same logic. Then (41) is equivalent to

$$\begin{aligned} \arg\min_b J(b) &= \int_{a=X_{\max}}^{\infty} \frac{|b-a|}{ab}\frac{1}{a^{n+3/2}} da \\ &= \int_{a=X_{\max}}^{b} \frac{b-a}{ab}\frac{da}{a^{n+3/2}} + \int_b^{\infty} \frac{a-b}{ab}\frac{da}{a^{n+3/2}} \\ &= \frac{2}{(n+1/2)(n+3/2)b^{n+3/2}} - \frac{1}{b(n+\frac{1}{2})X_{\max}^{n+\frac{1}{2}}} \\ &\quad + \frac{1}{(n+3/2)X_{\max}^{n+3/2}}. \end{aligned}$$

The minimum is found by setting the first derivative to zero:

$$\begin{aligned} J'(\hat{b}) &= \frac{2}{(n+1/2)(n+3/2)}\frac{(n+3/2)}{\hat{b}^{n+5/2}} \\ &\quad + \frac{1}{\hat{b}^2(n+1/2)X_{\max}^{n+1/2}} = 0 \\ \Rightarrow \hat{b} &= 2^{\frac{1}{n+1/2}} X_{\max}. \end{aligned}$$

To establish that $\hat{b}$ is in fact a minimum, note that

$$J''(\hat{b}) = \frac{1}{\hat{b}X_{\max}^{n+1/2}} = \frac{1}{2^{\frac{3}{n+1/6}} X_{\max}^{n+7/2}} > 0.$$

Thus, the restricted Bayesian estimate is the uniform distribution over $[0, 2^{\frac{1}{n+1/2}} X_{\max}]$.

REFERENCES

[1] B. Taskar, S. Lacoste-Julien, and M. I. Jordan, "Structured prediction, dual extragradient and Bregman projections," *Journal of Machine Learning Research*, vol. 7, pp. 1627–1653, 2006.

[2] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-Boost and Bregman divergence," *Neural Computation*, vol. 16, pp. 1437–1481, 2004.

[3] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, AdaBoost and Bregman distances," *Machine Learning*, vol. 48, pp. 253–285, 2002.

[4] J. Kivinen and M. Warmuth, "Relative loss bounds for multidimensional regression problems," *Machine Learning*, vol. 45, no. 3, pp. 301–329, 2001.

[5] J. Lafferty, "Additive models, boosting, and inference for generalized divergences," *Proc. of Conf. on Learning Theory (COLT)*, 1999.

[6] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Duality and auxiliary functions for bregman distances," *Carnegie Mellon University Technical Report CMU-CS-01-109R*, 2001.

[7] L. K. Jones and C. L. Byrne, "General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis," *IEEE Trans. on Information Theory*, vol. 36, pp. 23–30, 1990.

[8] M. R. Gupta, S. Srivastava, and L. Cazzanti, "Optimal estimation for nearest neighbor classifiers," *Univ. of Washington Dept. of Electrical Engineering Technical Report 2006-0006*, 2006, available at idl.ee.washington.edu.

[9] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1287–1314, 2007.

[10] A. Banerjee, "An analysis of logistic models: Exponential family connections and online performance," *SIAM Intl. Conf. on Data Mining*, 2007.

[11] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[12] R. Nock and F. Nielsen, "On weighting clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1223–1235, 2006.

[13] G. LeBesenerais, J. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. on Information Theory*, vol. 45, pp. 1565–1577, 1999.

[14] Y. Altun and A. Smola, "Unifying divergence minimization and statistical inference via convex duality," *Proc. of Conf. on Learning Theory (COLT)*, 2006.

[15] M. C. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem of information theory," *IEEE Trans. on Information Theory*, vol. 43, no. 4, pp. 1288–1293, 1997.

[16] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Trans. on Information Theory*, vol. 51, no. 7, pp. 2664–2669, 2005.

[17] I. Csiszár, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol. 68, pp. 161–185, 1995.

[18] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "An introduction to functional derivatives," *Univ. of Washington Dept. of Electrical Enginering Technical Report 2008-0001*, Available at idl.ee.washington.edu/publications.php.

[19] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. USA: Dover, 2000.

[20] T. Rockafellar, "Integrals which are convex functionals," *Pacific Journal of Mathematics*, vol. 24, no. 3, pp. 525–539, 1968.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[22] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer, 1998.

[23] S. Srivastava and M. R. Gupta, "Distribution-based Bayesian minimum expected risk for discriminant analysis," *Proc. of the IEEE Intl. Symposium on Information Theory*, 2006.

[24] T. Cover and J. Thomas, *Elements of Information Theory*. United States of America: John Wiley and Sons, 1991.

[25] R. E. Kass, "The geometry of asymptotic inference," *Statistical Science*, vol. 4, no. 3, pp. 188–234, 1989.

[26] S. Amari and H. Nagaoka, *Methods of Information Geometry*. New York: Oxford University Press, 2000.

[27] G. Lebanon, "Axiomatic geometry of conditional models," *IEEE Trans. on Information Theory*, vol. 51, no. 4, pp. 1283–1294, 2005.

[28] C. H. Edwards, *Advanced Calculus of Several Variables*. New York: Dover, 1995.